Hindawi
Complexity
Volume 2019, Article ID 9643921, 12 pages
https://doi.org/10.1155/2019/9643921

WILEY | Hindawi

Research Article

Outlier Detection and Correction for Monitoring Data of Water Quality Based on Improved VMD and LSSVM

Guangpei Sun,[1] Peng Jiang,[1] Huan Xu,[1] Shanen Yu,[1] Dong Guo,[2] Guang Lin,[3] and Hui Wu[4]

[1]*College of Automation, Hangzhou Dianzi University, Hangzhou 310018, China*
[2]*College of Electrical Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China*
[3]*Zhejiang Provincial Environmental Monitoring Center, Hangzhou 310018, China*
[4]*Fuzhou Fuguang Water Technology Co., Ltd., Fuzhou 350000, China*

Correspondence should be addressed to Peng Jiang; pjiang@hdu.edu.cn

Received 29 August 2018; Revised 12 November 2018; Accepted 4 December 2018; Published 3 February 2019

Academic Editor: Matilde Santos

To improve the detection rate and reduce the correction error of abnormal data for water quality, an outlier detection and correction method is proposed based on the improved Variational Mode Decomposition (improved VMD) and Least Square Support Vector Machine (LSSVM) algorithms. The correlation coefficient is introduced, for solving the optimal parameter k of VMD algorithm, and an improved VMD algorithm is obtained. Combined with LSSVM algorithm, the outliers of water quality can be detected and repaired. This method is applied for the detection and correction of water quality monitoring outliers using dissolved oxygen which is retrieved from the water quality monitoring station in Hangzhou, Zhejiang Province, China. The result shows that the improved VMD algorithm is of higher detection rate and lower error rate than those of Empirical Mode Decomposition (EMD) and Ensemble Empirical Mode Decomposition (EEMD). The LSSVM algorithm increases the fitting accuracy and decreases correction error in comparison with SVM and BP neural network, which provides important references for the implementation of environmental protection measures.

1. Introduction

Water resource is an important strategic resource of the country, and it has important influence on economic and social development. In recent years, people's awareness of environmental protection has been gradually strengthened, and the state's supervision of water pollution is gradually increasing. Water quality monitoring has become an important link in the process of water pollution control [1, 2]. Whether the water quality monitoring data is normal or not has a significant impact on the implementation of water environmental protection measures. Therefore, it is of great significance to detect and correct the outlier of monitoring data on water quality [3, 4].

Carlson and Byer applied the Pauta criterion to outlier detection of water quality for the first time, and it is assumed that data exceeding three sigma of the sample mean is outlier [5]. This method is simple and intuitive, and it can detect partial outlier. But the monitoring data of water quality has the characteristics of high monitoring frequency and large fluctuation; hence, this method is of bad accuracy in outlier detection of water quality [6]. On the basis of the above research, many scholars had conducted in-depth research on the outlier detection of water quality according to the water quality characteristics. Park S and his team used principal component analysis (PCA) to build a model [7], and outlier for monitoring data of water quality can be detected by calculating the residual error of the model. The PCA model reduced input dimension; however, the analysis results are poor in the case of low correlation between indicators. Hou D B's team took the monitoring value of ammonia nitrogen as an example [8] and detected the water quality data to be normal or not based on the wavelet transform and wavelet neural network (RBF); the model improves the detection rate and reduces the error rate to some extent. Each of these methods has common limitations, the detection

effect is poor when the data fluctuates greatly, and new methods [9–11] had been applied to abnormal data detection by scholars. Using the EMD algorithm to detect outlier for water quality, an anomaly detection method based on scale adaptive matching was proposed by Yang Z L [12]. The water quality anomaly detection is transferred to the time and frequency domain, and it provides a new idea for water quality outlier detection. However, the EMD algorithm [13, 14] has the modal mixing problem in the process of signal decomposition and the overall detection effect is affected. Zhang F and his team optimized the outlier detection method for water quality monitoring based on the EEMD algorithm [15], and ensemble empirical modal decomposition is applied to analyze abnormal monitoring data and reduce the problem of modal mixing effect, but the detection rate of abnormal data needs further improvement.

Through the analysis of previous studies, it is found that the detection of outlier for water quality has made great progress, but there is a need to improve detection rate although few scholars have corrected outlier detection. On the basis of these studies, this paper proposes an outlier detection and correction method for monitoring data of water quality based on improved VMD and LSSVM. This method has a higher detection rate and lower error rate than the EMD and EEMD algorithms. In addition, the paper adds to the correction of outlier by the LSSVM algorithm. Compared with SVM and BP neural network, LSSVM algorithm improves the fitting accuracy, and the error of reconstructing data is smaller. Finally, the algorithm package of this paper is useful for engineering application through our independently developed software platform.

2. Experimental Methods

2.1. The VMD Algorithm. The Variational Mode Decomposition (VMD) algorithm [16–18] is a variational problem solving process based on the classical Wiener filtering [19, 20], Hilbert transforms [21, 22], and frequency mixing, and it mainly includes two parts: the construction and solution of variational problems. The goal of VMD is to decompose an input signal into a discrete subsignal. Suppose that the signal f can be decomposed into a modal function $u_k(t)$ with the minimum sum of k bandwidth, and each mode has a central frequency with limited bandwidth, then the constraint condition is the sum of each mode equal to the input signal f. A variational model with constraints is established [23]:

$$\min_{\{u_k\},\{w_k\}} \quad \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\} \tag{1}$$

$$\boldsymbol{s.t.} \quad \sum_k u_k = f \tag{2}$$

where $\delta(t)$ is the Dirac distribution, $*$ represents convolution, u_k is the modal components, and w_k is the central frequency of the modal components.

To find optimal solution for the constrained variational model, it needs to be converted into a nonconstrained variational problem, and the secondary punitive factor α and the Lagrange multipliers λ are introduced.

$$\begin{aligned} &L(\{u_k\},\{w_k\},\lambda) \\ &= \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \\ &+ \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \left[\lambda(t), f(t) - \sum_k u_k(t) \right] \end{aligned} \tag{3}$$

The VMD algorithm adopts the alternate direction method of multipliers (ADMM) to solve the optimal solution of variational problem by alternating update parameters u_k^{n+1},w_k^{n+1}, and λ_k^{n+1} [24]. The solution is as follows:

(1) Initialize $\{u_k^1\}$, $\{w_k^1\}$, $\{\lambda_k^1\}$.

(2) Alternate update parameters u_k^{n+1}, w_k^{n+1} and λ_k^{n+1}.

$$u_k^{n+1} = \underset{u_k}{\operatorname{argmin}} \, \mathrm{L}\left(\left\{ u_{i<k}^{n+1} \right\}, \left\{ u_{i\geq k}^{n} \right\}, \left\{ w_i^n \right\}, \lambda^n \right) \tag{4}$$

$$w_k^{n+1} = \underset{w_k}{\operatorname{argmin}} \, \mathrm{L}\left(\left\{ u_i^{n+1} \right\}, \left\{ w_{i<k}^{n+1} \right\}, \left\{ w_{i\geq k}^{n} \right\}, \lambda^n \right) \tag{5}$$

$$\lambda^{n+1} = \lambda^n + \tau \left(f - \sum_k u_k^{n+1} \right) \tag{6}$$

(3) Repeat step (2), when

$$\frac{\sum_k \left\| u_k^{n+1} - u_k^n \right\|_2^2}{\left\| u_k^n \right\|_2^2} < \varepsilon \quad \left(\varepsilon = 1 \times 10^{-6} \right) \tag{7}$$

Stop updating, and get k number of intrinsic mode function (IMF) [17].

2.2. The LSSVM Algorithm. The LSSVM (Least Square Support Vector Machine) algorithm is an improved algorithm for SVM (Support Vector Machines). LSSVM is a statistical learning theory that adopts a least squares linear system as a loss function [25] and it transforms inequality constraints into equality constraints, takes the loss function of the sum of squared errors as the empirical loss of training set, and turns the empirical risk from first power to second power. Finally, the quadratic programming problem is transformed into linear equations and is solved by least square method. The speed of solution and the accuracy of convergence are both improved [26].

Let us say there is a nonlinear sample set S, and $S = \{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$, the training sample x_i and y_i ($x_i \in R^n$, $y_i \in$ R) is an n-dimensional vector, and we map the original space sample from R^n to the feature space $\varphi(x_i)$ by a nonlinear mapping $\varphi(\bullet)$. The optimal decision function is constructed in this high dimensional feature space:

$$y(x) = \boldsymbol{\omega} \varphi(x) + b \tag{8}$$

Use structural risk minimization principle to solve parameters $\boldsymbol{\omega}$, b. The problem of function fitting becomes the following optimization problem:

$$\min \quad \boldsymbol{J}(\boldsymbol{\omega}, e) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2 \tag{9}$$

$$\boldsymbol{s.t.} \quad y_i = \boldsymbol{\omega}^T\varphi(x_i) + b + e_i, \quad i = 1, 2, \ldots, n \tag{10}$$

In the formula: $\boldsymbol{\omega}$ is the weight vector, γ is the regularization parameter, e_i is the error variable, and b is the deviation. Using the Lagrangian method to solve this optimization problem:

$$\boldsymbol{L}(\boldsymbol{\omega}, b, e, a) = \boldsymbol{J}(\boldsymbol{\omega}, e) - \sum_{i=1}^{n} a_i\left[\boldsymbol{\omega}^T\varphi(x_i) + b + e_i - y_i\right] \tag{11}$$

In this formula: a_i is the Lagrange multiplier.

According to Kuhn-Tucher conditions [27]:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\omega}} &= 0\\ \frac{\partial L}{\partial b} &= 0\\ \frac{\partial L}{\partial e_i} &= 0\\ \frac{\partial L}{\partial a_i} &= 0\end{aligned} \tag{12}$$

The following formulae can be obtained:

$$\begin{aligned}\boldsymbol{\omega} &= \sum_{i=1}^{n} a_i\varphi(x_i)\\ \sum_{i=1}^{n} a_i &= 0\\ a_i &= \gamma e_i\\ \boldsymbol{\omega}^T\varphi(x_i) + b + e_i - y_i &= 0\end{aligned} \tag{13}$$

Selecting the radial basis function:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\delta^2}\right) \tag{14}$$

In this formula: δ is the kernel parameter and $K(x, x_i)$ is the kernel function. The optimization problem is transformed into solving a linear algebraic system of equations:

$$\begin{bmatrix} 0 & \boldsymbol{I}_n^{\mathrm{T}} \\ \boldsymbol{I}_n & \varphi(x_i)^{\mathrm{T}}\varphi(x_i) + \left(\frac{1}{\gamma}\right)\boldsymbol{I} \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{15}$$

In (15), $\boldsymbol{I}$ is a column vector.

$$\begin{aligned}\boldsymbol{I}_n &= [1, 1, \ldots, 1]^{\mathrm{T}}\\ a &= [a_1, a_2, \ldots, a_n]^{\mathrm{T}}\\ y &= [y_1, y_2, \ldots, y_n]^{\mathrm{T}}\end{aligned} \tag{16}$$

Using least square method to solve a and b, the linearized expression has been achieved:

$$f(x) = \sum_{i=1}^{n} a_i K(x, x_i) + b \tag{17}$$

3. Improved Experimental Method

3.1. The Improved VMD Algorithm. The VMD algorithm searches for the optimal solution of the variational model by nonrecursive iteration in frequency domain and determines the center frequency and bandwidth of each amplitude modulation and frequency modulation component, and finally adaptive frequency division and separation of components are realized [28]. This method has high resolution and can effectively avoid the problem of pattern confusion, but the precondition of obtaining the optimal decomposition results is that the number k of modal decomposition is determined in advance [29]. Traditionally, empirical method or referring EMD mode decomposition is used to determine the k value of parameters; in order to overcome the difficulty of solving the optimal value of parameter k in VMD algorithm, an improved VMD algorithm based on Newton method is proposed. The specific steps are as follows:

(1) Construct objective function: the correlation coefficient (COR) is introduced, which represents the degree of correlation between the modal decomposition signal and the original signal [30]. When the COR value is below the threshold, it is considered that the modal decomposition margin no longer contains information related to the original signal, and the original signal is completely decomposed. The correlation coefficient $\rho_{\mathrm{xy}}(k)$ of two time series $x_k(n)$ and $y(n)$ is defined as follows:

$$\rho_{\mathrm{xy}}(k) = \frac{\sum_{n=0}^{\infty} x_k(n)\, y(n)}{\sqrt{\sum_{n=0}^{\infty} x_k^2(n) \sum_{n=0}^{\infty} y^2(n)}} \tag{18}$$

Among them, $x_k(n)$ is the k modal component, $y(n)$ is the original signal, and $\rho_{\mathrm{xy}}(k)$ is the correlation coefficient between the k mode component and the original signal. The objective function is defined as follows:

$$f(k) = \rho_{\mathrm{xy}}(k) \tag{19}$$

(2) Iterative search optimal solution [31]: set the initial iteration parameter k=1, execute the iteration loop according to k=k+1, and stop iterating until $f(k) \leq \varepsilon$ (in this paper, ε takes 0.2). The algorithm block diagram for solving the optimal parameter k by Newton's method is shown in Figure 1.

(3) The k value is introduced into the VMD algorithm to get the optimal mode decomposition.

3.2. The Optimization of LSSVM Algorithm. In the LSSVM algorithm, the regularization parameter γ and the kernel width δ are two very important parameters. If the γ is too small, it will lead to less-learning and otherwise it can lead to over-learning. The value of δ directly affects the accuracy of

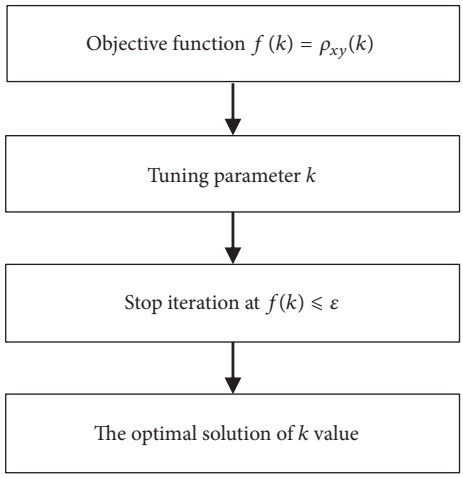


Figure 1: The algorithm block diagram of Newton's method.

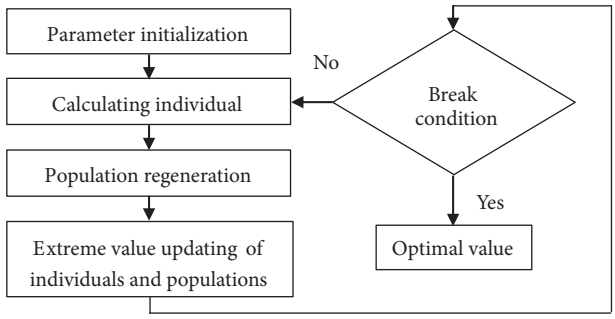


Figure 2: The flow chart of LSSVM algorithm optimizations.

model fitting [32]. In this paper, particle swarm optimization (PSO) algorithm is used to optimize these two parameters to improve the performance of LSSVM in curve fitting. The flow chart of LSSVM algorithm optimization is shown in Figure 2.

The specific steps are as follows [33]:

(1) Parameter initialization: initialization of PSO parameters, including population size, learning factors, and inertia weight.

(2) Calculating individual fitness: the fitness value of each particle is calculated by LSSVM model, and then the current fitness value is compared with the best fitness value of the particle itself to get the optimal position of the particle.

(3) Population regeneration: comparing the optimal position fitness of each particle with the optimal position fitness of the population, the best one is selected as the optimal position of the population, and the position and velocity of each particle in the population are updated.

(4) When the number of iterations reaches the maximum, the optimization is finished and the current optimal particle is selected as the parameter of LSSVM algorithm. Otherwise, jump to step (2), and continue to do iterative optimization.

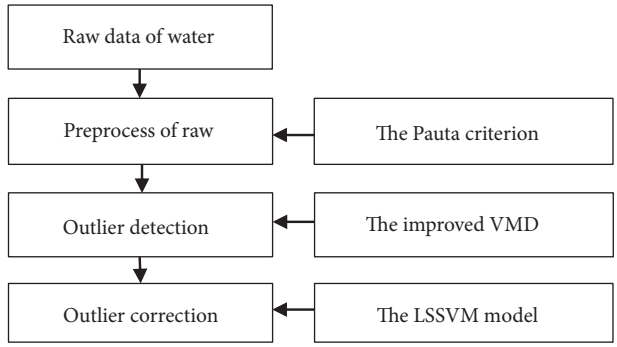


Figure 3: The block diagram of water quality outlier detection and correction model.

3.3. The Model of Outlier Detection and Correction for Water Quality. In order to improve the detection rate and reduce the correction error of abnormal data for water quality, an outlier detection and correction method for monitoring data of water quality was proposed based on improved VMD and LSSVM. First of all, the data of water quality monitoring station is preprocessed by the rule of the Pauta criterion to eliminate the obvious outlier. Then, the improved VMD algorithm is used in the mode decomposition for the residual monitoring data sequence and the outlier of monitoring data was detected by superimposing the low frequency modal components. Finally, the LSSVM algorithm is used to correct the outlier. The detailed block diagram of water quality outlier detection and correction model is shown in Figure 3.

4. Results and Discussion

4.1. Performance Comparison of Signals Decomposition. Using the EMD and improved VMD algorithms to decompose the simulation signals and to compare the performance of the two algorithms in signal decomposition, the simulation signals are composed of three cosine signals of 55Hz, 266Hz, and 580Hz and a group of noise sequence A; the expression of the simulation signal is

$$f(t) = \cos(2 * pi * 55 * t) + \frac{1}{2} * \cos(2 * pi * 266 * t) + \frac{1}{4} * \cos(2 * pi * 580 * t) + A \quad (20)$$

The expression of A is as follows:

A = zeros(1,1024)

$A([8, 16, 25, 30, 67, 69, 78, 97, 101, 134, 150, 170, 210, 245, 289, 310, 330, 400, 420, 440, 506]) = 0.6$

$A([536, 562, 581, 602, 635, 665, 693, 726, 742, 771, 800, 825, 847, 862, 879, 893, 1005]) = -0.6$

In this experiment, the sampling frequency is 5120Hz and the sampling number is 1024. The time domain graph and the corresponding spectrum graph of the simulation signal $f(t)$ are shown in Figures 4 and 5, respectively.

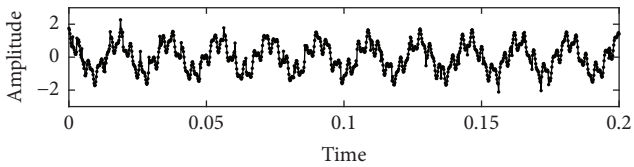


Figure 4: The time domain graph of the simulation signal.

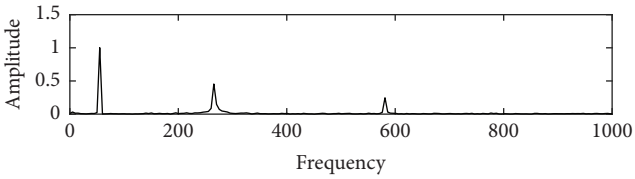


Figure 5: The corresponding spectrum graph of the simulation signal.

As can be seen from Figure 5, the simulation signal mainly contains the frequency components of 55Hz, 266Hz, and 580Hz. Using the improved VMD algorithm to decompose the original signal, the secondary penalty factor α and the number of modal decomposition k need be determined first.

For the selection of α, the value of α will affect the decomposition effect of the improved VMD algorithm. The smaller the α, the greater the bandwidth of the intrinsic mode function (IMF) components and, conversely, the less bandwidth of the IMF's components. According to experience, the value of α is usually 5000~10000. Furthermore, the change of α in the appropriate range will not have too big impact on the decomposition effect. In this experiment, the value of α is 8000.

For the selection of k values, if the k value is too large, the IMF components decomposed will be intermittent and the optimal solution k=4 is obtained by iteration of Newton method. The decomposition results and corresponding spectra by improved VMD are shown in Figures 6 and 7, respectively.

The EMD algorithm is used to decompose the simulation signals, and the decomposition results and corresponding spectra are shown in Figures 8 and 9, respectively.

From Figures 6 and 7 we can see that four IMF signal components are decomposed through the improved VMD algorithm. The frequency of the first three IMF signal components is 55.1076Hz, 265.5186Hz, and 581.135Hz, respectively, and consistent basically with the frequency components contained in the original signal. The fourth IMF component is a set of noise signals with very low intensity, mainly distributed in the frequency range from 1500Hz to 2100Hz. As can be seen from Figures 8 and 9, the decomposition result is not ideal through the EMD algorithm. The phenomenon of modal overlap occurs in IMF components from the third to the fifth. As the experimental results suggest, the decomposition effect of EMD algorithm is not ideal and caused mode mixing problem. The improved VMD algorithm can overcome the disadvantage of mode mixing problem and achieve good decomposition effect. To sum up, the improved

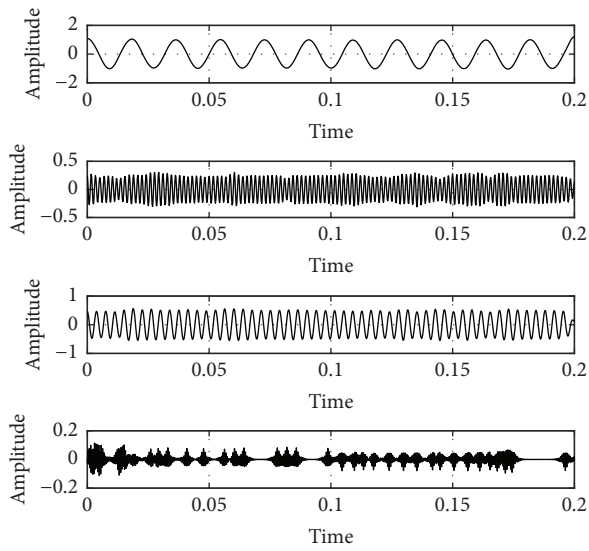


Figure 6: Results of signal decomposition by improved VMD.

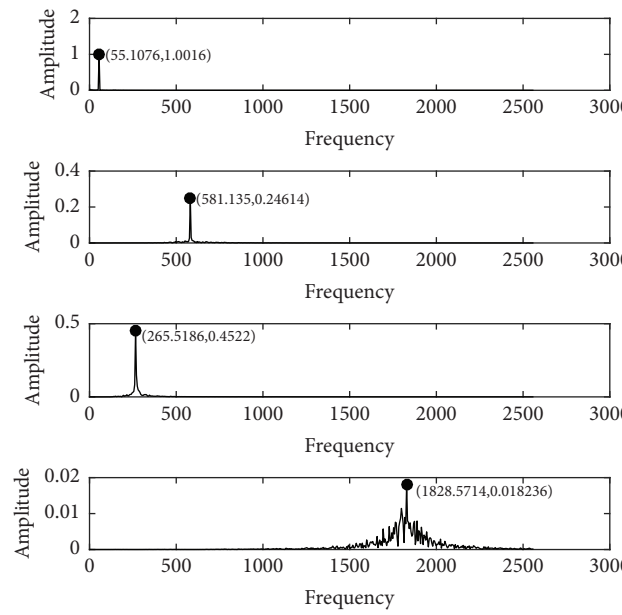


Figure 7: Relevant spectra figures by improved VMD.

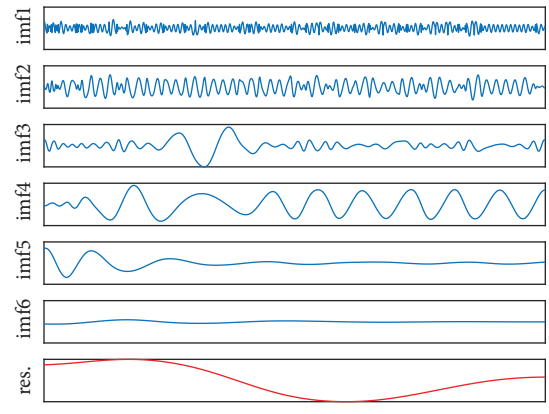


Figure 8: Results of signal decomposition by EMD.

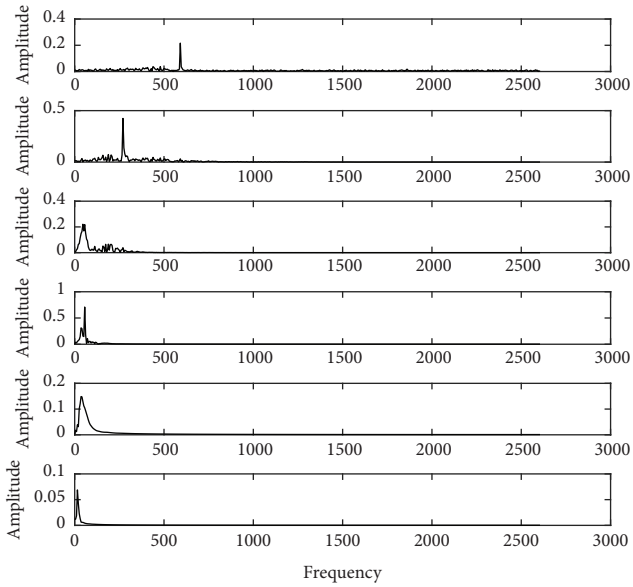


Figure 9: Relevant spectra figures by EMD.

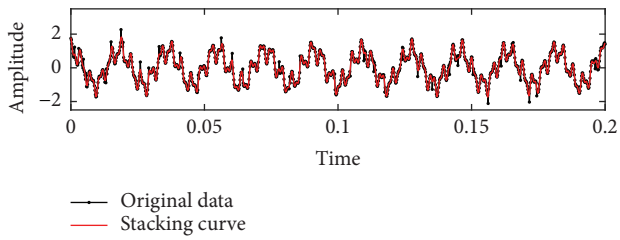


Figure 10: The time domain diagram of original signal and stacking signal.

VMD algorithm is better than the EMD algorithm for signal decomposition.

4.2. Outlier Detection and Correction. The fourth IMF component in Figure 6 is a series of extremely low noise signals, which needs to be removed. In order to detect the abnormal points in the original signal, add the remaining three IMF components in Figure 6 and a new time series signal $f'(t)$ is obtained. The time domain diagram of $f'(t)$ and $f(t)$ is shown in Figure 10.

Taking ±50% as the threshold of relative error between original data sequence and the stacking data is calculated. The data is treated as outlier when the relative error exceeds threshold. In Figure 11, the spot marked with red dots is the outlier detected in the original simulation signal.

In order to verify the superiority of improved VMD algorithm in outlier detection, two comparative experiments were designed: using the EMD algorithm and EEMD algorithm to detect outliers. The detection results of the two algorithms are shown in Figures 12 and 13, respectively.

The detection results of EMD, EEMD, and improved VMD are shown in Table 1.

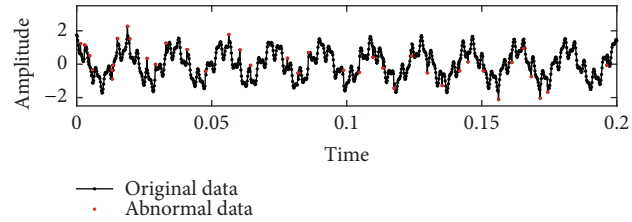


Figure 11: Outlier detection results by improved VMD for simulated data.

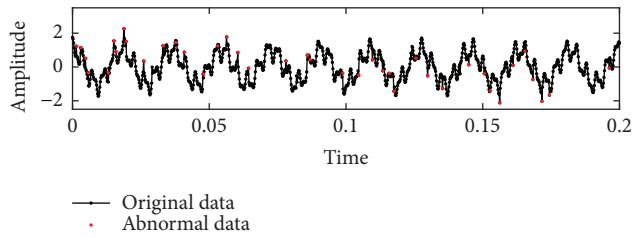


Figure 12: Outlier detection results by EMD for simulated data.

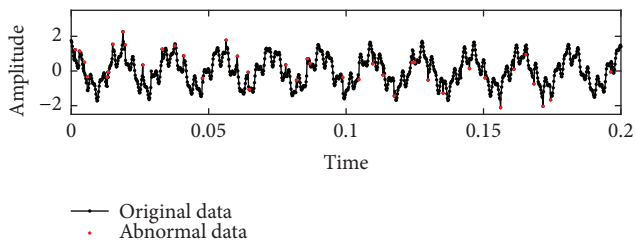


Figure 13: Outlier detection results by EEMD for simulated data.

In order to calculate the accuracy of three algorithms for outlier detection, the number of normal data is distinguished into normal data labeled as TP, the number of abnormal data is distinguished into normal data labeled as FP, the number of normal data is distinguished into abnormal data labeled as FN, and the number of abnormal data is distinguished into abnormal data labeled as TN. The calculation methods for the detection rate Acc(Accuracy) and the error rate Fal(False) of the abnormal data are shown by the following formula.

$$Acc = \frac{TN}{(FP + TN)} \times 100\% \quad (21)$$

$$Fal = \frac{FN}{(TP + FN)} \times 100\% \quad (22)$$

According to formula (21) and formula (22), the detection rate and error rate of EMD, EEMD, and improved VMD algorithms are shown in Table 2.

As seen in Table 2, in the aspect of outlier detection, the accuracy of EMD algorithm is the lowest in the three algorithms, which is 86.84%. The modal mixing in the decomposition process of EMD algorithm is an important reason leading to low accuracy. In the same way, the EMD algorithm has a higher error rate, which is 0.71%. The EEMD algorithm has a promotion compared to the EMD algorithm, and it can be seen that the detection rate of EEMD has

Table 1: Detection results of abnormal data.

Element	Number of sampling	Number of outlier	Number of detected	Number of false detected
EMD	1024	38	33	7
EEMD	1024	38	36	4
improved VMD	1024	38	37	2

Table 2: Detection rate and error rate.

Element	TP	FP	FN	TN	Acc (%)	Fal (%)
EMD	979	5	7	33	86.84	0.71
EEMD	982	2	4	36	94.74	0.41
improved VMD	984	1	2	37	97.37	0.20

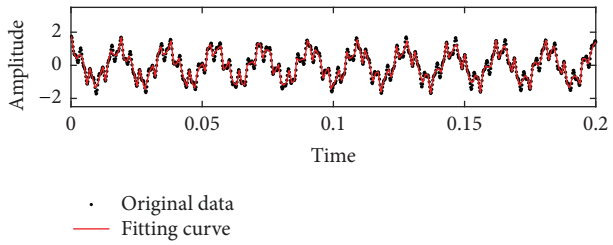


Figure 14: The fitting result by LSSVM.

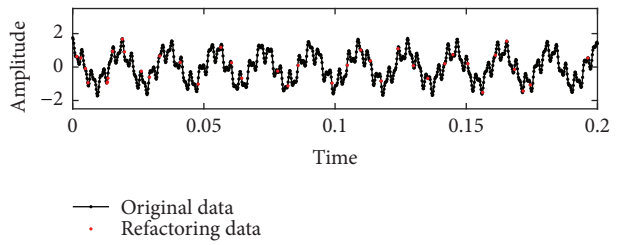


Figure 15: The correction result of outlier.

been greatly improved, which is 94.74%, and the error rate is also greatly reduced to 0.41%. The improved VMD algorithm proposed in this paper has obvious advantages in signal decomposition precision and noise robustness, it also can be seen that the detection rate of improved VMD algorithm for abnormal data is further increased, which is 97.37%, and the error rate is the lowest of the three algorithms, which is 0.20%. In the three algorithms, the improved VMD algorithm has the best effect for outlier detection.

Remove the abnormal data detected in Figure 10, and the remaining sampling points constitute a new set of sequences. The dispersion normalization method is used to process the sequence.

$$f(t) = \frac{(f(t) - f_{min})}{(f_{max} - f_{min})} \tag{23}$$

The normalized fitting result is $f_0(t)$, and the actual fitting result is as follows.

$$f = f_0(t) \times (f_{max} - f_{min}) + f_{min} \tag{24}$$

The parameters γ and δ of the LSSVM model are determined by the PSO method, and take the final result $\gamma = 703$, $\delta^2 = 1.8$ as the optimal parameter. The curve is fitted with LSSVM model, and the result is shown in Figure 14.

The correction result of outlier is shown in Figure 15.

In order to increase the contrast of the experiment, two algorithms of SVM and BP neural network are used to correct the outlier detected, respectively. The value of correcting outlier by the LSSVM, SVM, and BP neural network algorithms is shown in Figure 16.

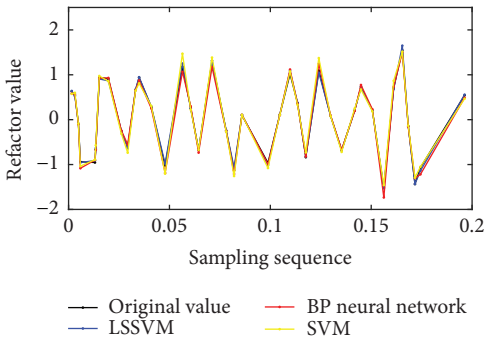


Figure 16: The value of correcting outlier by three algorithms.

MSE, MAE, and MAPE indicators are used to evaluate the performance of the algorithm for outlier correction, and the results are shown in Table 3.

According to the experimental results of Table 3, the fitting effect of SVM algorithm is the worst of the three algorithms, and the reason is that the selection of kernel function in SVM algorithm is difficult. Compared with the SVM algorithm, the BP neural network algorithm has improved the correction effect for outlier, but it is not obvious because the algorithm is dependent on the selection of training set samples. The LSSVM algorithm adopted in this paper is the improvement of SVM algorithm, and the effect of data fitting is obviously improved. In addition, the value of the MAPE index of the three algorithms is obviously larger than the two indexes of MSE and MAE, which is because the value of the data set selected in this experiment is small and does not affect the performance evaluation of the algorithm. Taken

Table 3: Comparison results of the algorithm performance.

Element	SVM	BP neural network	LSSVM
MSE	0.0098	0.0061	0.0013
MAE	0.0722	0.0556	0.0260
MAPE	10.319	7.8877	3.4540

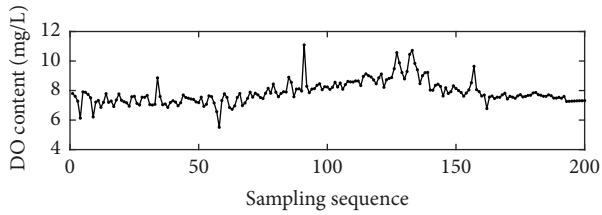


Figure 17: The DO concentration of the monitoring site.

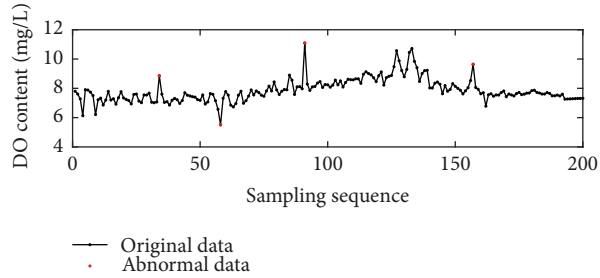


Figure 18: The preliminary outlier detection results.

together, the LSSVM algorithm has the best fitting effect; that is, LSSVM has the highest accuracy in outlier correction.

4.3. Outlier Detection and Correction of Actual Monitoring Data. Take the monitoring data of DO for a period of time in a water quality monitoring station in Hangzhou (Wan Cun station from Jan 1, 2018 to Feb 2, 2018) as an example, and record it as $x(t)$. The time sequence for the DO concentration of this monitoring site is shown in Figure 17.

In order to simplify the operation of improved VMD algorithm, the raw data of DO is preprocessed first. According to the Pauta criterion in classical statistics, the preliminary outlier detection results are shown in Figure 18.

After preliminary detection, four outliers are detected. In Figure 18, the sample points marked with red dots are outlier. Removing the four obvious outliers, a new set of sampling sequences is obtained. Using the improved VMD algorithm to decompose the sampling sequence, the optimal mode decomposition number k=3 is obtained by Newton method. The parameters in the improved VMD algorithm are set as follows: the value of α is 8000, and the value of k is 3. The time domain graph and the corresponding spectrum graph of the mode decomposition are shown in Figure 19.

Remove the third IMF component and superpose the remaining two IMF components. A superimposed sequence is obtained, as shown in Figure 20.

Selecting the threshold according to the method in the simulation experiment, the outlier is detected and shown in Figure 21.

As shown in Figure 21, 9 outliers are detected through the VMD algorithm. In addition to the 4 outliers detected during pretreatment, 13 outliers in the DO monitoring value are detected in this experiment. Using the LSSVM algorithm to correct the 13 abnormal data, the correction results are shown in Figure 22.

In order to verify the effectiveness of this method in practical engineering application, we add a set of comparative experiments additionally. A set of standard monitoring data was obtained from Zhejiang Provincial Environmental Protection Department. The data was 200 samples of DO content at Jiu Xi monitoring station from April 1, 2018 to May 3, 2018. Then 20 samples were artificially modified to represent abnormal samples. EMD, EEMD, and improved VMD algorithm is used respectively to detect the outliers, and LSSVM algorithm is used to repair the outliers. The results are shown in Figures 23(a)–23(c) and Table 4, respectively.

The results show that the improved VMD algorithm is of great effect on outlier detection and it has high accuracy and low error rate. From the two indicators of detection rate and error rate, we can see that the performance of the improved VMD algorithm is better than that of EMD and EEMD algorithms, which is consistent with the experimental results in simulated data scenarios.

For the outlier correction, the comparison of algorithm performance among SVM, BP neural network, and LSSVM is shown in Table 5.

The result in Table 5 is also consistent with the experimental results in simulated data scenarios. For the MSE,

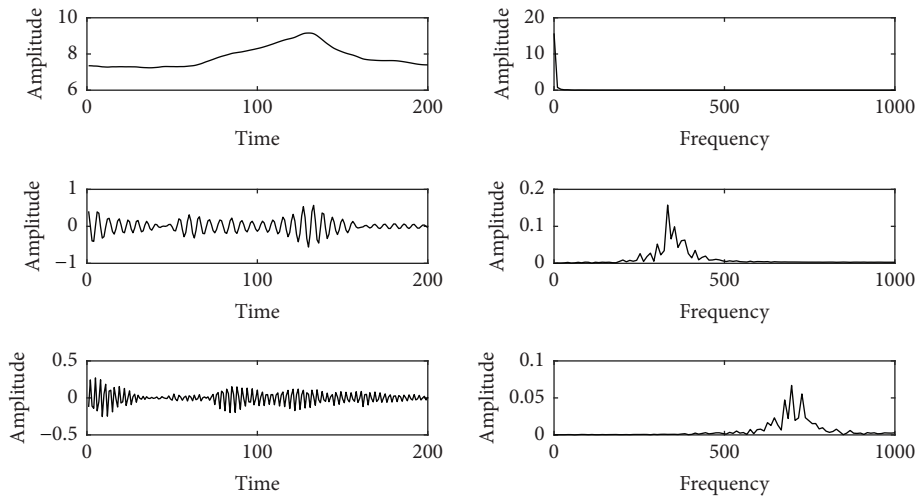


Figure 19: Decomposition result and corresponding spectrum by improved VMD.

Table 4: Detection rate and error rate for actual data.

Element	TP	FP	FN	TN	Acc (%)	Fal (%)
EMD	178	4	2	16	82.00	1.11
EEMD	179	2	1	18	90.00	0.56
improved VMD	179	1	1	19	95.00	0.56

Table 5: Comparison of the algorithm performance for actual data.

Element	SVM	BP neural network	LSSVM
MSE	0.0108	0.0073	0.0021
MAE	0.0785	0.0571	0.0280
MAPE	4.704	3.1630	1.6667

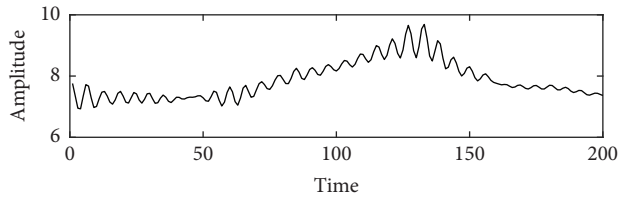


Figure 20: The superimposed sequence.

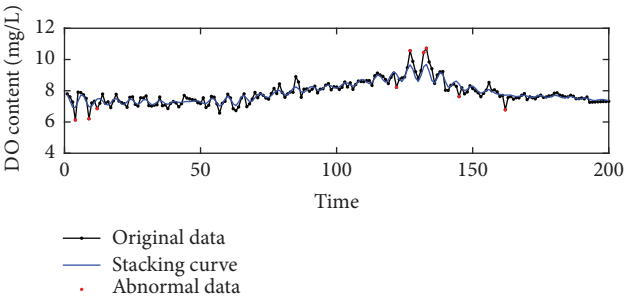


Figure 21: Outlier detection results for DO monitoring value.

MAE, and MAPE three indicators, the performance of the LSSVM algorithm is obviously better than that of SVM and BP neural network.

The method of this paper has already realized the engineering application of algorithm package in the water quality parameters monitoring and trend forecast system developed by ourselves, and it has been applied to water quality monitoring stations in Zhejiang Province, China. This method avoids the data error caused by equipment failure, external interference, and other factors. It also substitutes traditional artificial statistics and correction and improves the efficiency and service quality of environmental protection. The location of water quality stations in the software platform is shown in Figure 24.

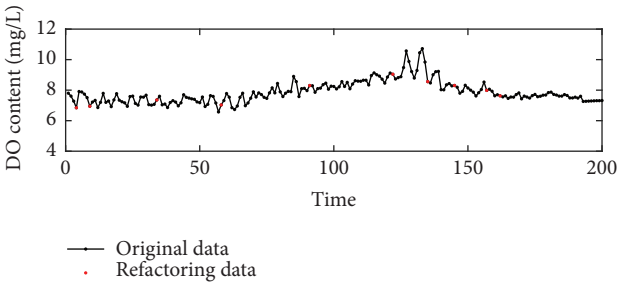
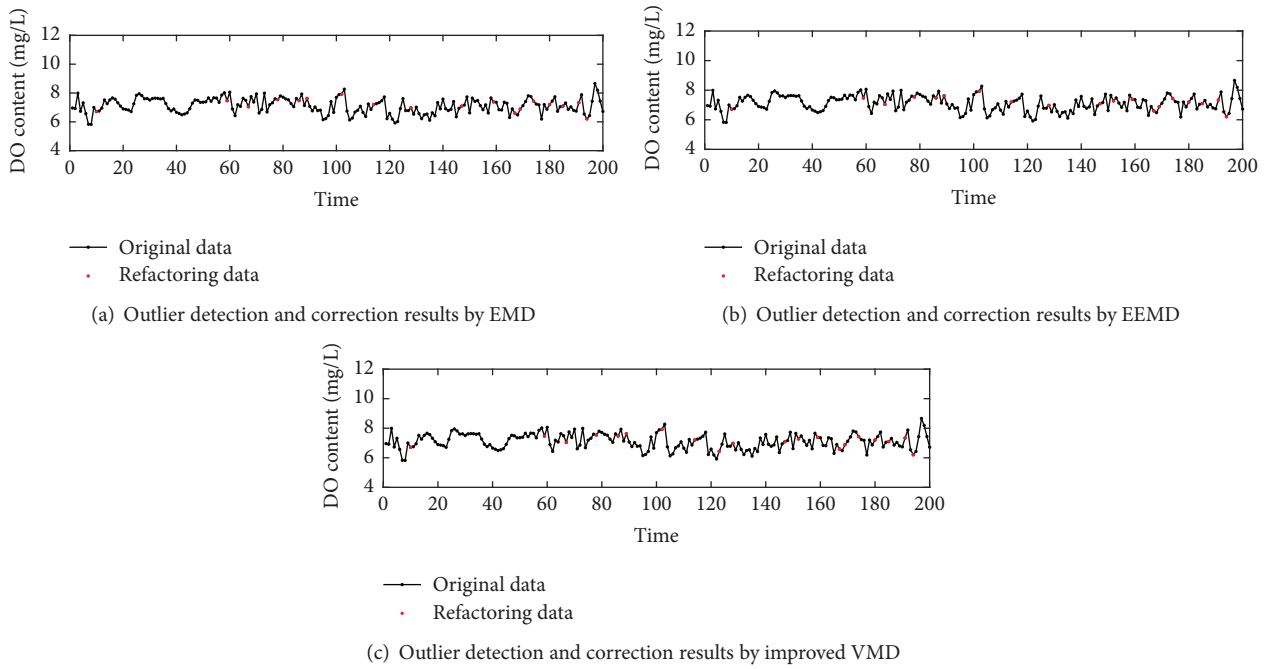


Figure 22: Outlier correction results for DO monitoring value.



(a) Outlier detection and correction results by EMD

(b) Outlier detection and correction results by EEMD

(c) Outlier detection and correction results by improved VMD

Figure 23

Figure 24: The location of water quality monitoring stations.

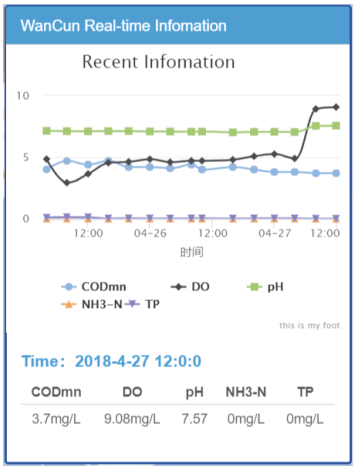


Figure 25: The engineering implementation effect of algorithm package.

Using the method of this paper, abnormal values of DO concentration in different monitoring stations are detected and corrected. Taking the Wan Cun monitoring site of Hangzhou as an example, the engineering implementation effect of the algorithm package is shown in Figure 25. The black graph shows the historical curve of the DO concentration value of the monitoring station processed through the algorithm package.

In future, the algorithm will be applied to more water quality parameters, such as COD, PH, NH_3-N, and TP, which will be more practical.

5. Conclusions

To improve the detection rate and reduce the error rate of outlier for water quality data, an outlier detection and correction method based on improved VMD and LSSVM is proposed, and the method is applied to Wan Cun which is a water quality monitoring station in Hangzhou. The method avoids the shortcoming of EMD algorithm in the process of signal decomposition. On the indicator of detection rate and error rate, the method of this paper is superior to the algorithm of EMD and EEMD. Based on the outlier detection, the outlier of DO monitoring value is corrected. On the indicator of MSE, MAE, and MAPE, improved VMD is better than the algorithm of SVM and BP neural network. The method proposed in this paper can be applied to water quality monitoring and its related fields, which will provide

an important reference for the enforcement of environmental department and the implementation of environmental protection measures.

Data Availability

The real-time monitoring data used in the manuscript were obtained from the Drinking Water Quality Automatic Monitoring Station of Zhejiang Environmental Protection Department collected from Jan/01/2018 to May/31/2018. Any researcher can see http://yys.zjemc.org.cn/Home/Map?moudleName=realtime# for more information.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by International Science and Technology Cooperation Program of Zhejiang Province for Joint Research in High-Tech Industry (No. 2016C54007), National Natural Science Foundation of China and Zhejiang Joint Fund for Integrating of Informatization and Industrialization (No. U1509217), Provincial Key R&D Program of Zhejiang Province (No. 2017C03019), and the National Key R&D Program of China (No. 2016YFC0201400).

Supplementary Materials

A graphical summary of the manuscript that let readers quickly capture the core content of the paper. *(Supplementary Materials)*

References

[1] C. Yang, "National Water Resources Monitoring Capacity Building Project and its Progress," *Water Resources Informatization*, 2013.

[2] Y. Chen, K. Zhao, Y. Wu et al., "Spatio-Temporal Patterns and Source Identification of Water Pollution in Lake Taihu (China)," *Water*, vol. 8, no. 3, p. 86, 2016.

[3] Y. Jiang and Y. Wan, "Demand analysis of water resources monitoring capacity and implementation strategy study," *China Water Resources*, 2012.

[4] T. M. Huang, J. G. Zhai, R. Wang et al., "The Detection of Abnormal Data in Marine Survey," *Acta Geodaetica Et Cartographic Sinica*, vol. 107, no. 1270, pp. 67–70, 2015.

[5] D. Byer and K. H. Carlson, "Expanded Summary: Real-time detection of intentional chemical contamination in the distribution system," *Journal - American Water Works Association*, vol. 97, no. 7, pp. 130–133, 2005.

[6] H. Fang, H. Xue, Y. Jiang, T. Zhou, Y. Wan, and H. Wang, "Outlier Detection and Correction for Water Resources Monitoring Data Based on EEMD," *Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery*, vol. 48, no. 9, pp. 257–263, 2017.

[7] S. Park and S.-Y. Jung, "Principal component analysis of water pipe flow data," *Procedia Engineering*, vol. 89, pp. 395–400, 2014.

[8] B. D. Hou, Y. Chen, F. H. Zhao et al., "Based on the RBF neural network and wavelet analysis the water quality of anomaly detection method," *Transducer and Microsystem Technologies*, vol. 32, no. 2, pp. 138–141, 2013.

[9] C. Chen, X. Lin, and G. Terejanu, "An Approximate Bayesian Long Short-Term Memory Algorithm for Outliers Detection," 2017.

[10] A. M. Garay, H. Bolfarine, V. H. Lachos, and C. R. Cabral, "Bayesian analysis of censored linear regression models with scale mixtures of normal distributions," *Journal of Applied Statistics*, vol. 42, no. 12, pp. 2694–2714, 2015.

[11] D. G. Eliades, T. P. Lambrou, C. G. Panayiotou, M. M. Polycarpou et al., "Contamination event detection in water distribution systems using a model-based approach," *Procedia Engineering*, vol. 89, pp. 1089–1096, 2014.

[12] Z. L. Yang, *Based on Empirical Mode Decomposition of Urban Water Supply Water Quality Abnormal Event Detection Method Research*, Zhejiang university, 2016.

[13] G. Rilling and P. Flandrin, "One or two frequencies? The empirical mode decomposition answers," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 85–95, 2008.

[14] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis* , vol. 30, no. 2, pp. 243–261, 2011.

[15] F. Zhang, X. Huifeng, W. Wang et al., "Water resources monitoring abnormal data mode decomposition - support vector machine reconstruction method," *Journal of Agricultural Machinery*, vol. 48, no. 11, pp. 316–323, 2017.

[16] M. Feldman, *Decomposition and Analysis of Non-Stationary Dynamic Signals Using the Hilbert Transform*, American Society of Mechanical Engineers, 2008.

[17] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[18] Y. Wang, R. Markert, J. Xiang, and W. Zheng, "Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system," *Mechanical Systems and Signal Processing*, vol. 60, pp. 243–251, 2015.

[19] S. M. LaConte, S.-C. Ngan, and X. Hu, "Wavelet transform-based Wiener filtering of event-related fMRI data," *Magnetic Resonance in Medicine Official Journal of the Society of Magnetic Resonance in Medicine*, vol. 44, no. 5, pp. 746–757, 2000.

[20] D. K. Ramanah, G. Lavaux, and B. D. Wandelt, "Wiener filter reloaded: fast signal reconstruction without preconditioning," *Monthly Notices of the Royal Astronomical Society*, vol. 468, no. 2, pp. 1782–1793, 2017.

[21] H. Huang and J. Pan, "Speech pitch determination based on Hilbert-Huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006.

[22] Z.-X. Hu and W.-X. Ren, "Vibration signal demodulation and instantaneous frequency estimation based on recursive Hilbert transformation," *Journal of Vibration & Shock*, vol. 35, no. 7, pp. 39–43, 2016.

[23] Q. Xiao, J. Li, Z. Bai, J. Sun, N. Zhou, and Z. Zeng, "A small leak detection method based on VMD adaptive de-noising and ambiguity correlation classification intended for natural gas pipelines," *Sensors*, vol. 16, no. 12, 2016.

[24] T. Erseghe, "Distributed optimal power flow using ADMM," *IEEE Transactions on Power Systems*, vol. 29, no. 5, pp. 2370–2380, 2014.

[25] P. Samui and D. P. Kothari, "Utilization of a least square support vector machine (LSSVM) for slope stability analysis," *Scientia Iranica*, vol. 18, no. 1, pp. 53–58, 2011.

[26] W. Li, Y. P. Yang, and N. Wang, "Multi-model LSSVM regression modeling based on kernel fuzzy clustering," *Kongzhi yu Juece/Control and Decision*, vol. 23, no. 5, pp. 560–562, 566, 2008.

[27] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[28] S.-K. Liu, G.-J. Tang, and X.-L. Wang, "Time frequency analysis method for rotary mechanical fault based on improved variational mode decomposition," *Journal of Vibration Engineering*, vol. 29, no. 6, pp. 1119–1126, 2016.

[29] Y. Yue, G. Sun, Y Cai et al., "Application of variational mode decomposition in bearing fault diagnosis," *Bearings*, vol. 8, pp. 50–54, 2016.

[30] J. Lee, "Correction: a note on the concordance correlation coefficient," *Biometrics*, vol. 56, no. 1, pp. 324-325, 2000.

[31] Y. Huihui, W. Yongli, C. Yongyong et al., "Distributed quasi Newton algorithm for solving unconstrained conformance optimization problems," *Journal of Shandong University of Science and Technology (NATURAL SCIENCE EDITION)*, vol. 35, no. 3, pp. 112–118, 2016.

[32] W. Mei, *An Improved Parameter Selection Method for Nuclear Function*, Xi'an University of Science And Technology, 2011.

[33] X. Wu, Q. Huang, and X. J. Zhu, "Thermal modeling of a solid oxide fuel cell and micro gas turbine hybrid power system based on modified LS-SVM," *International Journal of Hydrogen Energy*, vol. 36, no. 1, pp. 885–892, 2011.